·论著·一次研究·

面向肺纤维化关键基因的Stacking集成分析

张　旭[1]，张石萍[2]，郝智航[1]，李俊燕[2]，谈娜娜[2]，倪士峰[3]，王　乐[2]，胡静波[4]，王　欢[1]

1. 宝鸡文理学院计算机学院（陕西宝鸡 721000）
2. 宝鸡文理学院化学与材料工程学院（陕西宝鸡 721013）
3. 西北大学生命科学学院（西安 710069）
4. 宝鸡文理学院电子电气工程学院（陕西宝鸡 721000）

【摘要】目的 构建基于Stacking策略的集成学习模型，以提升高维小样本肺纤维化（PF）基因表达数据的特征筛选稳定性，并识别与疾病相关的候选关键基因。**方法** 从基因表达综合数据库（GEO）数据库获取PF转录组数据集GSE70866与GSE48149，其中GSE70866作为训练集用于模型构建与特征筛选，GSE48149作为独立验证数据集用于评估模型的泛化能力，经数据预处理与差异表达分析筛选候选关键基因。在此基础上构建多个基学习器的Stacking集成框架，通过K折交叉验证生成元特征，并以逻辑回归（LR）和支持向量机（SVM）作为元学习器完成最终分类判别。采用F1-score与曲线下面积（AUC）评价模型性能，结合特征重要性排序筛选候选基因。**结果** 构建的集成学习模型在训练数据集GSE70866中表现出良好的判别能力，F1-score为0.955 9，AUC为0.948 2，整体优于单一模型方法。在独立验证数据集GSE48149上，混淆矩阵结果进一步表明该模型具有较好的泛化能力。基于多模型特征重要性综合分析，筛选获得硫氧还蛋白样蛋白4B（*TXNL4B*）、C-C基序趋化因子配体18（*CCL18*）、泛素结合酶E2 Z亚型（*UBE2Z*）等候选关键基因。**结论** 基于Stacking策略的集成学习框架可在高维小样本条件下提升模型稳定性与预测准确性，为转录组数据中的候选基因筛选提供了一种有效方法，为PF相关分子过程的探索提供数据层面支持。

【关键词】肺纤维化；Stacking集成；逻辑回归；支持向量机；特征重要性
【中图分类号】R966　　**【文献标识码】**A

Stacking ensemble analysis of key genes in pulmonary fibrosis

ZHANG Xu[1], ZHANG Shiping[2], HAO Zhihang[1], LI Junyan[2], TAN Nana[2], NI Shifeng[3], WANG Le[2], HU Jingbo[4], WANG Huan[1]

1. School of Computer, Baoji University of Arts and Sciences, Baoji 721000, Shaanxi Province, China
2. College of Chemistry and Material Engineering, Baoji University of Arts and Sciences, Baoji 721013, Shaanxi Province, China
3. College of Life Sciences, Northwestern University, Xi'an 710069, China
4. School of Electronic and Electrical Engineering, Baoji University of Arts and Sciences, Baoji 721000, Shaanxi Province, China
Corresponding author: WANG Huan, Email: hwang227@126.com

DOI: 10.12173/j.issn.2097-4922.202603067
基金项目：国家自然科学基金青年科学基金项目（82104682）；陕西省科技厅项目（2025JC-YBQN-1249、2015JM8463）；陕西省科技创新团队（2022TD-63）；宝鸡文理学院校级研究生创新科研项目（YJSCX25YB65）
通信作者：王欢，博士，教授，硕士研究生导师，Email：hwang227@126.com

【Abstract】Objective To construct an ensemble learning model based on the stacking strategy, and to improve the stability of feature screening for high-dimensional, small-sample pulmonary fibrosis (PF) gene expression data and to identify candidate key genes associated with the disease. Methods The PF transcriptome datasets GSE70866 and GSE48149 were obtained from the Gene Expression Omnibus (GEO) database. GSE70866 was used as the training set for model construction and feature selection, while GSE48149 was used as an independent validation dataset to evaluate the model's generalization ability. Candidate key genes were screened after data preprocessing and differential expression analysis. In this study, a stacking ensemble framework composed of multiple base learners was constructed. Meta-features were generated through K-fold cross-validation, and Logistic regression (LR) together with support vector machine (SVM) were employed as Meta-learners for final classification. Model performance was evaluated using the F1-score and the area under the curve (AUC), while candidate genes were identified based on feature importance ranking. Results The constructed ensemble learning model demonstrated strong discriminative performance on the training dataset GSE70866, achieving an F1-score of 0.955 9 and an AUC of 0.948 2, outperforming individual models overall. On the independent validation dataset GSE48149, the confusion matrix further indicated that the model possessed good generalization capability. Based on integrated feature importance analysis across multiple models, candidate key genes were identified, including thioredoxin-like protein 4B (*TXNL4B*), C-C motif chemokine ligand 18 (*CCL18*), and ubiquitin-conjugating enzyme E2 Z (*UBE2Z*). Conclusion The stacking-based ensemble learning framework can improve model stability and prediction accuracy under high-dimensional, small-sample conditions, providing an effective method for screening candidate genes in transcriptome data and offering data-level support for exploring PF-related molecular processes.

【Keywords】Pulmonary fibrosis; Stacking ensemble; Logistic regression; Support vector machine; Feature importance

肺纤维化（pulmonary fibrosis，PF）是一类以肺组织结构持续性重塑和呼吸功能进行性退化为特征的慢性致命性疾病，其发生发展涉及上皮–间质转化、免疫微环境失衡、氧化应激紊乱以及细胞外基质（extracellular matrix，ECM）异常沉积等多种信号通路的协同失调。近年来，PF已成为多种结缔组织病相关并发症中导致患者死亡的主要原因之一，其起病隐匿、进展迅速及预后不良等特点，严重影响患者生存质量并增加临床治疗负担。目前临床应用的抗纤维化药物，如尼达尼布和吡非尼酮，虽可在一定程度上延缓疾病进展，但难以逆转已形成的纤维化结构，且不同患者之间治疗响应差异明显。PF的分子调控网络尚未完全阐明，系统识别驱动疾病发生发展的关键调控基因，对于深入解析PF复杂分子机制、发现早期诊断标志物及探索潜在治疗靶点具有重要意义[1]。

随着高通量测序技术的发展，基因表达谱数据为疾病分子机制研究提供了重要支撑。传统的差异表达基因（differentially expressed genes，DEGs）及加权基因共表达网络分析在候选基因筛选方面发挥了一定作用，但在处理高维、小样本组学数据时，往往难以充分刻画基因间复杂的非线性关系与交互效应，筛选结果的稳定性和泛化能力受到一定限制[2]。近年来，机器学习方法在生物标志物筛选与疾病分类研究中得到广泛应用。支持向量机（support vector machine，SVM）、随机森林（random forest，RF）等模型能够在一定程度上提升预测性能，但单一模型在高维、强噪声数据环境中易受特征冗余及样本分布偏倚影响，泛化能力有限。集成学习通过融合多模型预测结果，可综合不同模型的优势，提高模型稳定性与鲁棒性[3]。其中，Stacking框架通过构建多层学习结构整合基学习器输出，为复杂生物数据分析提供了一种有效策略。

基于此，本研究构建了一种基于Stacking策略的集成学习模型，对PF转录组数据进行系统分析。通过整合多种基学习器的判别结果，筛选潜在关键调控基因，并结合功能富集与信号通路分析，对候选关键基因的生物学作用进行解析[4]。基因组学多层数据流图见图1。本研究旨在提高

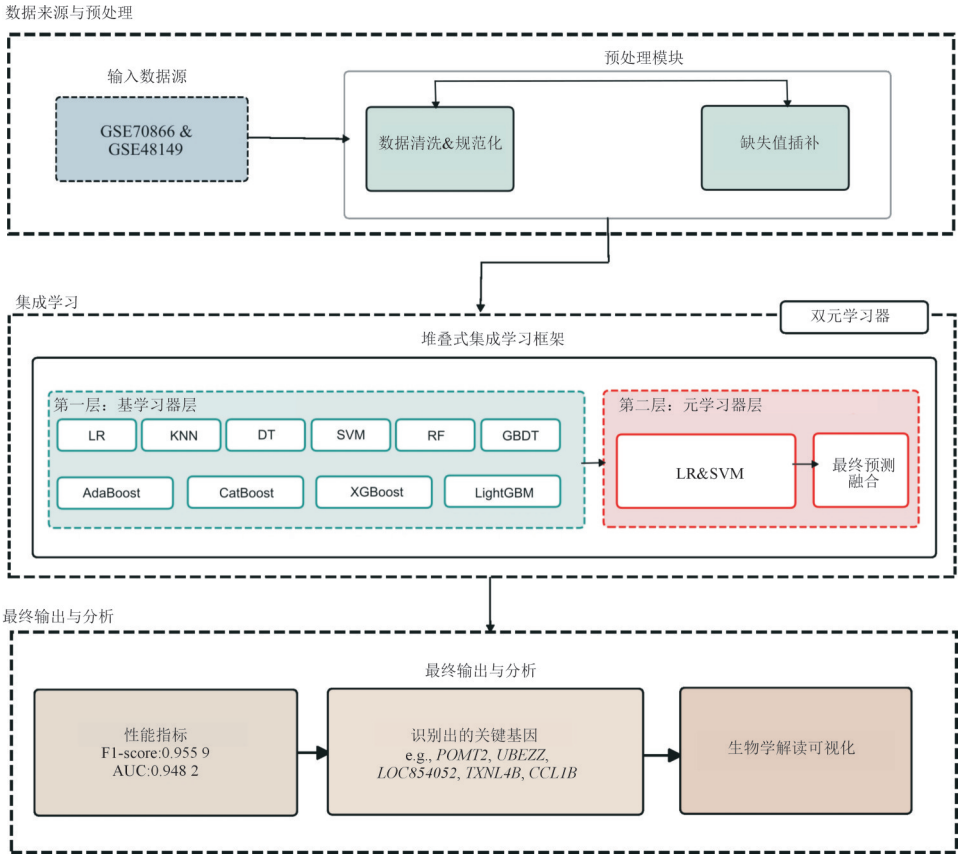

图1 基于Stacking集成学习的PF基因组学多层数据流图

Figure 1. Multilevel data flow diagram of PF genomics based on stacking ensemble learning

注：GEO. Gene Expression Omnibus；LR. 逻辑回归；KNN. K近邻；DT. 决策树；SVM. 支持向量机；RF. 随机森林；GBDT. 梯度提升决策树；AdaBoost. 自适应增强；CatBoost. 类别特征梯度提升；XGBoost. 极端梯度提升；LightGBM. 轻量梯度提升机；AUC. 曲线下面积；POMT2. 蛋白质O-甘露糖基转移酶2；UBEZZ. 泛素结合酶Z；LOC854052. 未表征基因LOC854052；TXNL4B. 硫氧还蛋白样4B；CCL1B. 趋化因子配体1B。

高维小样本数据分析的稳定性与可靠性，为PF分子机制研究及个体化靶向干预策略的探索提供技术路径与理论依据。

1 资料与方法

1.1 数据来源与预处理

本研究选取基因表达综合数据库（Gene Expression Omnibus，GEO）中公开的PF相关数据集GSE70866和GSE48149为研究对象[5-6]。以“pulmonary fibrosis”为关键词，利用R软件包GEOquery检索PF相关数据集GSE70866（176例PF患者和20名健康对照样本）和GSE48149（8例PF患者和27名健康对照样本），在后续模型训练、混淆矩阵分析及性能评价中，将PF定义为正类（positive class，label=1），健康对照定义为负类（negative class，label=0）。由于样本规模及实验设计存在差异，本研究将GSE70866作为训练数据集，将GSE48149作为独立验证数据集，以评估筛选关键基因的稳定性和泛化能力。

为保证实验设计的一致性与可重复性，本研究的数据划分策略如下：在训练阶段，基于GSE70866数据集采用分层5折交叉验证进行模型训练与参数优化；在验证阶段，使用未参与训练的GSE48149数据集作为独立验证数据集评估模型性能。所有模型训练、调参与评估过程均严格限定在训练数据范围内，避免信息泄露。数据预处理阶段，首先对原始芯片数据进行背景校正与标准化处理。针对表达矩阵中极少量缺失值（缺失比例低于1%），采用基因维度均值填补策略，即以该基因在其他样本中的平均表达值进行填补。为减小不同基因表达量级差异对模型训练的影响，对表达数据进行$\log_2$转换，并进一步采用Z-score标准化方法进行归一化处理，以提高后续特征筛选与分类分析的稳定性。经上述预处理后，获得标准化的基因表达矩阵，进行后续差异表达分析及集成学习模型构建。

1.2 基于Stacking的集成判别

为有效融合多种基分类器的判别能力，针对小样本数据过拟合问题，本研究采用Stacking集成学习框架，通过分层集成策略实现预测结果的协同优化，该框架由基学习器层与元学习器层共同构成。

1.2.1 第一层：基学习器层

在该层中，基因表达矩阵被分别输入至10类基学习器。每个基学习器独立完成训练，并在交叉验证机制下生成对训练数据集GSE70866的预测输出。为避免信息泄漏并提高模型泛化能力，本研究在Stacking框架中采用基于折外（out-of-fold，OOF）预测策略的分层K折交叉验证策略生成元特征。具体而言，将训练数据划分为K=5个互斥子集。在每一折中，基学习器仅在其中K-1折数据上进行训练，并对剩余1折未参与训练的数据进行预测，从而获得对应样本的预测输出。通过遍历所有折次，确保每个样本的预测结果均来自未见数据（OOF预测），最终拼接形成完整的元特征矩阵。该策略有效避免了在同一数据上同时进行训练与预测所带来的信息泄漏问题，从而保证元学习器输入特征的独立性与可靠性，该设置在样本量有限的生物医学数据分析中，能够在偏差与方差之间实现较优平衡。在超参数调优过程中，本研究采用网格搜索（grid search）方法，在预设参数空间内对各基学习器进行穷举搜索。具体参数范围设定如下：SVM中惩罚参数$C \in \{0.1, 1, 10\}$，核函数参数*gamma* $\in$ {scale，auto}；RF中树的数量*n_estimators* $\in$ {100，200，300}，最大深度*max_depth* $\in$ {3，5，10}，叶节点最小样本数 *min_samples_leaf* $\in$ {1，3，5}；梯度提升决策树（gradient boosting decision tree，GBDT）、极端梯度提升算法（extreme gradient boosting，XGBoost）等梯度提升类模型中学习率*learning_rate* $\in$ {0.01，0.05，0.1}，基学习器数量*n_estimators* $\in$ {100，300}，最大深度*max_depth* $\in$ {3，4，6}，子采样比例*subsample* $\in$ {0.8，1.0}；K近邻（K-nearest neighbors，KNN）中邻居数*n_neighbors* $\in$ {3，5，7}等。所有参数组合均在5折分层交叉验证框架下进行评估，以F1-score作为优化指标，其定义为精确率与召回率的调和平均值，最终选择性能最优的参数组合[7]，其最优配置汇总见表1。鉴于本研究数据集样本规模较小，若进一步采用嵌套交叉验证可能导致训练数据进一步划分，从而增加模型方差并影响稳定性。因此，本研究在分层K折交叉验证框架下完成参数优化与模型评估，在保证计算效率的同时兼顾结果的稳健性。针对超参数最终，所有基学习器的预测结果被横向拼接，构成一个高维的元特征矩阵，作为第二层的输入。对应计算公式如下：

$$h_m(X) = \hat{y}_m,\quad m = 1, 2, \cdots, M \qquad \text{公式（1）}$$

$$Z = \left[\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_M\right] \qquad \text{公式（2）}$$

公式（1）表示第m个基学习器$h_m(X)$对输入特征X的预测输出，其中$m = 1, 2, \cdots, M$。公式（2）表示将所有基学习器的预测结果进行横向拼接，构成元特征向量Z，其维度为M。该元特征矩阵作为第二层元学习器的输入，实现模型集成。

1.2.2 第二层：元学习器层

元学习器仅基于上述OOF预测生成的元特征进行训练，学习如何最优地组合各基模型的预测结果，从而做出最终判别确保整个Stacking模型

表1 不同机器学习模型的最优超参数设置

Table 1. Optimal hyperparameter settings for different machine learning models

模型	超参数设置
LR	C=0.01；max_iter=1 000；penalty=l2；solver=lbfgs
KNN	n_neighbors=5；weights=distance
DT	max_depth=5；min_samples_leaf=5
SVM	C=1.0；gamma=scale；kernel=rbf
RF	n_estimators=100；max_depth=5；min_samples_leaf=5
GBDT	n_estimators=100；max_depth=5；min_samples_leaf=5
AdaBoost	n_estimators=100；learning_rate=0.01
CatBoost	iterations=300；learning_rate=0.05；depth=6
XGBoost	n_estimators=300；learning_rate=0.05；max_depth=4；subsample=0.8
LightGBM	n_estimators=300；learning_rate=0.05；max_depth=4；subsample=0.8

注：LR：逻辑回归；DT：决策树；AdaBoost：自适应提升算法；CatBoost：类别提升算法；LightGBM：轻量级梯度提升机。

在训练过程中不存在数据泄漏风险。本研究采用LR与SVM作为双元学习器，构建并行堆叠式集成模型，2种元学习器并行接收相同的元特征并独立完成训练与预测，在最终预测阶段，LR与SVM分别输出正类概率。本文采用概率平均策略对两者输出结果进行融合，即对2个元学习器预测概率取均值作为最终预测结果。

在第二层学习阶段，LR作为概率型判别模型，用于学习元特征与类别标签之间的线性决策关系[8]。其输出正类概率定义为：

$$P(y=1|x)=\sigma\left(w^{T}x|b\right) \quad \text{公式（3）}$$

其中，Z表示由第一层基学习器生成的元特征向量，w为权重向量，b为偏置项，$\sigma(\cdot)$为Sigmoid激活函数。Sigmoid函数定义为：

$$\sigma(z)=\frac{1}{1+e^{-z}} \quad \text{公式（4）}$$

该函数将线性组合结果映射到区间（0，1），从而实现对类别概率的建模。为优化模型参数，采用二元交叉熵损失函数进行训练：

$$L=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log\left(\hat{y}_i\right)+\left(1|y_i\right)\log\left(1|\hat{y}_i\right)\right] \quad \text{公式（5）}$$

其中，N为样本数量，y_i为真实标签，$\hat{y}_i$为模型预测概率。该损失函数等价于最大化样本的对数似然函数，能够有效度量预测概率与真实标签之间的偏差。在堆叠结构中，LR具有参数少、可解释性强和对小样本数据适应性良好的优势，因此适合作为稳定的概率型元学习器。

SVM是一种基于结构风险最小化原则的判别模型，通过最大化分类间隔以增强模型泛化能力[9]。其优化目标函数为：

$$\min_{w,\ b}\frac{1}{2}\ //\ w\ //\ ^{2}+C\sum_{i=1}^{N}\xi_i \quad \text{公式（6）}$$

其中，$//\ w\ //\ ^{2}$控制分类间隔大小，ξ_i为松弛变量，用于允许部分样本违反间隔约束，C为惩罚系数，用于平衡间隔最大化与误分类惩罚之间的权衡。其约束条件为：

$$y_i\left(w^{T}\phi\left(x_i\right)|b\right)\geqslant 1-\xi_i \quad \text{公式（7）}$$

其中，$\phi(\cdot)$表示核映射函数，用于将元特征映射至高维特征空间，从而实现非线性决策边界的学习。通过核函数机制，SVM能够捕获元特征之间的潜在非线性关系。在样本规模有限且特征维度较高的生物医学数据场景中，SVM具有较强的泛化能力与抗过拟合特性。

LR能够高效建模基学习器预测结果之间的线性加权关系，具有良好的可解释性与稳定性；SVM在引入核函数后，可有效捕捉基学习器输出间的非线性交互模式，在复杂边界判别任务中表现优异。最终预测阶段，LR与SVM分别输出正类概率，采用概率平均策略对两者输出结果进行融合，即对2个元学习器预测概率取均值作为最终预测结果。元学习器概率融合公式为：

$$P_{final}=\frac{1}{2}\left(P_{LR}|P_{SVM}\right) \quad \text{公式（8）}$$

其中，P_{LR}和P_{SVM}分别表示LR与SVM输出的正类概率。该策略能够在保证模型稳定性的同时，融合线性与非线性判别优势，提高最终预测性能。

综合考虑性能、效率与鲁棒性，本研究最终采用LR与SVM作为双元学习器方案，通过该层的融合判别机制，在保留各基学习器优势的同时，显著提升了整体模型的预测精度与泛化能力。

1.3 模型构建

为系统评估所构建Stacking模型的分类性能与关键基因筛选效果，本研究从内部验证、外部验证及生物学层面3个维度开展分析。其中，内部验证基于训练数据集GSE70866的分层5折交叉验证结果；外部验证基于独立验证数据集GSE48149评估模型在未见数据上的分类能力；生物学层面分析则结合候选关键基因在独立验证数据集GSE48149中的表达差异验证及KEGG富集分析，从表达模式与潜在作用通路2个方面对筛选结果进行综合支持。

1.3.1 基于GSE70866数据集的分层5折交叉验证

为评估所构建Stacking集成模型在训练数据集上的分类性能，本研究基于训练数据集GSE70866采用分层5折交叉验证策略，对各类模型进行系统比较。所有性能指标均以$\bar{x}\pm s$形式表示，以减少随机划分带来的波动影响。

在对比模型设置方面，单一模型包括传统机器学习模型与深度学习模型两类。其中，传统机器学习模型涵盖LR、KNN、DT、SVM、RF、GBDT、AdaBoost、XGBoost及LightGBM。同时，为进一步评估模型在复杂非线性特征建模方面的能力，引入卷积神经网络（convolutional neural

network，CNN）与Transformer模型作为深度学习对比方法。对于CNN模型，采用一维卷积神经网络结构（1D-CNN），以标准化后的基因表达向量作为输入，网络由卷积层、ReLU 激活函数及最大池化层构成，并通过全连接层输出分类结果。训练过程中采用 Adam 优化器，学习率设为0.001，批大小（batch size）为 16，训练轮数（epoch）为100。对于Transformer模型，构建基于自注意力机制的编码结构，将基因表达数据视为序列输入，通过多头注意力机制提取全局特征表示。模型训练同样采用Adam优化器，并引入早停（early stopping）策略以抑制过拟合风险。所有深度学习模型均采用与传统机器学习模型一致的数据划分策略，以确保实验结果具有良好的可比性。

1.3.2 基于独立验证数据集的模型评估

为进一步评估模型在未参与训练的数据上的分类能力，本研究选取独立验证数据集GSE48149，对训练完成的Stacking模型进行外部验证分析。为直观展示模型在样本层面的分类效果，进一步绘制混淆矩阵。混淆矩阵从样本层面对模型预测结果进行了直观展示，其中横轴表示预测类别，纵轴表示真实类别。

1.4 关键基因筛选

在完成模型训练后，本研究进一步整合多模型特征重要性评分与差异表达分析结果，以筛选高置信度候选关键基因。整体筛选流程如下：首先，在独立验证数据集GSE48149（基于 Illumina Human HT-12 V4.0 expression beadchip，GPL16221 平台）上进行DEGs分析。采用 limma 包构建线性模型，并通过 Benjamini-Hochberg 方法进行多重检验校正以控制假发现率（false discovery rate，FDR）。

在此基础上，结合集成模型中各基学习器的特征重要性评分，构建多模型协同筛选策略。具体而言：在每个基学习器中选取特征重要性排名前20%的基因作为候选集合，并定义“模型支持度”为某基因被选入的模型数量。仅保留支持度≥3的基因作为高一致性候选基因，以增强筛选结果的稳定性与鲁棒性。此外，为进一步验证筛选结果的稳定性，本研究进一步引入LASSO回归作为独立特征选择方法用于辅助验证分析。

本研究参考He等[10]关于肺相关疾病转录组分析的研究，为结果的合理性与可比性提供外部依据。采用UpSet图揭示不同模型筛选结果的交集关系，本研究对最终确定的核心基因在独立验证数据集GSE48149的病例组与对照组中进行表达差异验证。

1.5 京都基因与基因组百科全书信号通路富集分析

为进一步从信号通路层面解析筛选得到的候选关键基因在PF中的潜在作用机制，本研究对核心候选基因进行京都基因与基因组百科全书（Kyoto Encyclopedia of Genes and Genomes，KEGG）通路富集分析。

1.6 SHAP分析

为更细致地理解最终构建的Stacking集成模型在个别样本上的预测依据，本文引入SHAP（Shapley additive explanations）分析方法，对关键特征的贡献进行定量和可视化分析[11]。SHAP分析基于博弈论中的Shapley值理论，将每个特征视为合作博弈中的参与者，模型预测结果视为整体收益。通过计算特征在不同子集组合中的边际贡献，量化其对最终预测结果的影响。设特征集合为 F，i 为其中某一特征，SHAP 值定义为：

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

公式（9）

其中，F表示全部特征集合，S为不包含特征i的任意特征子集，$f(S)$表示仅基于特征子集S构建模型时的预测输出，$f(S \cup \{i\}) - f(S)$则表示特征i在给定子集S条件下对模型输出所产生的边际贡献。通过SHAP值，可以明确哪些特征对模型判别起到了积极或消极的作用，从而揭示模型的决策机制，与生物学理解相结合，增强模型的可信度。

2 结果

2.1 Stacking集成模型

基于训练数据集GSE70866的分层5折交叉验证，各类基学习器（如XGBoost、SVM、LightGBM 等）、CNN和Transformer作为单一模型时表现中等，本研究提出的模型在Recall、F1-score和曲线下面积（area under the curve，AUC）指标上均优于单一模型。结果表明，通过Stacking框架融合多种基学习器的预测结果，能够有效整合不同算法的优势，可提升模型的判别能力与鲁棒性。具体见表2。

表2 各单一模型与Stacking模型的性能比较（$\bar{x} \pm s$）

Table 2. Performance comparison of individual models and the stacking model ($\bar{x} \pm s$)

模型	Recall	F1-score	AUC
LR	0.818 2 ± 0.098	0.821 2 ± 0.098	0.866 7 ± 0.067
KNN	0.818 2 ± 0.098	0.818 2 ± 0.098	0.857 1 ± 0.014
DT	0.636 4 ± 0.016	0.642 4 ± 0.018	0.758 5 ± 0.017
SVM	0.909 1 ± 0.022	0.910 6 ± 0.014	0.900 0 ± 0.013
RF	0.727 3 ± 0.011	0.731 9 ± 0.015	0.857 1 ± 0.016
GBDT	0.690 3 ± 0.020	0.818 2 ± 0.021	0.821 4 ± 0.027
AdaBoost	0.909 1 ± 0.096	0.910 6 ± 0.053	0.816 7 ± 0.021
CatBoost	0.727 3 ± 0.014	0.731 9 ± 0.015	0.892 9 ± 0.022
XGBoost	0.760 0 ± 0.023	0.760 0 ± 0.023	0.821 4 ± 0.021
LightGBM	0.636 4 ± 0.060	0.694 9 ± 0.077	0.600 0 ± 0.011
CNN	0.714 3 ± 0.024	0.595 2 ± 0.013	0.800 0 ± 0.011
Transformer	0.833 3 ± 0.025	0.757 6 ± 0.025	0.909 5 ± 0.012
Stacking	0.954 5 ± 0.011	0.955 9 ± 0.009	0.948 2 ± 0.020

在独立验证数据集GSE48149上，Stacking模型在27例健康对照样本中均实现了正确分类，即真负例（true negative，TN）为27；在8例PF样本中，模型正确识别了7例，即真正例（true positive，TP）为7，仅有1例被误判为健康对照，即假负例（false negative，FN）为1。模型在独立验证数据集GSE48149上的Recall与F1-score分别为0.875和0.933，结果表明，本研究所构建的Stacking集成模型在保持较高灵敏度的同时，也具备良好的特异性，在未参与训练的数据上能够较好地区分PF患者与健康个体，具有较好的分类能力与泛化性能（图2）。

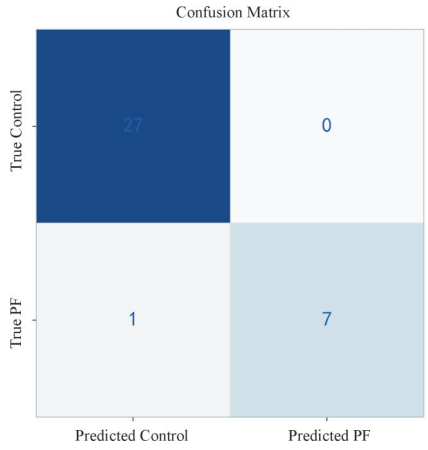

图2 GSE48149数据集上Stacking模型的混淆矩阵

Figure 2. Confusion matrix of the Stacking model on the GSE48149 dataset

注：label 1=PF；label 0=control。

2.2 关键基因

在严格筛选阈值下（$|log_2FC| > 1.5$且adj.$P < 0.05$），研究共鉴定DEGs 151个，其中上调78个，下调73个。结果显示，部分基因在不同模型中反复出现，呈现较高的一致性，例如*POMT2*（ILMN_21092）和*UBE2Z*（ILMN_17384）分别在5个模型中被选出，*LOC654052*（ILMN_39439）则在4个模型中出现。此外，*TXNL4B*、*RNPEPL1*、*PBXIP1*、*GALNTL2*、*CCL18*、*TP53*、*ZNF579*、*LTV1*、*VPS4A*、*PON2*、*UTY*、*ASNS*、*C14ORF143*、*TIMM9*等基因也在多个模型中呈现交集。

对不同模型的特征选择能力进行比较，结果显示，LR和AdaBoost各识别出最多的候选基因（均占41.2%），表明这两种模型在特征选择中具有较高的敏感性；XGBoost次之，占比35.3%，显示其在复杂特征模式捕捉上的优势。树模型类算法（如RF、GBDT、CatBoost、LightGBM）之间的交集较少，提示其在样本空间划分和特征权重计算上存在一定差异性；而LR与学习模型（AdaBoost、XGBoost）的交集较多，说明线性判别特征与加权集成特征存在一定重叠，增强了多模型协同筛选的鲁棒性（图3）。在满足“模型支持度≥3”的基础上，进一步结合差异表达显著性（$|log_2FC|$）及特征重要性综合排序，最终选取得分最高的5个基因作为核心候选关键基因。本研究筛选出的关键基因为*POMT2*、*UBE2Z*、*LOC654052*、*CCL18*、*TXNL4B*，候选关键基因功能与相关信息见表3。

ILMN_21092、ILMN_17384、ILMN_39439、ILMN_6746和ILMN_5441为对应的5个候选关键基因探针。候选关键基因在病例组与对照组之间存在明显表达差异，且总体变化趋势与训练集结果一致，部分候选关键基因在独立验证数据集GSE48149中仍表现出显著的表达差异（图4）。

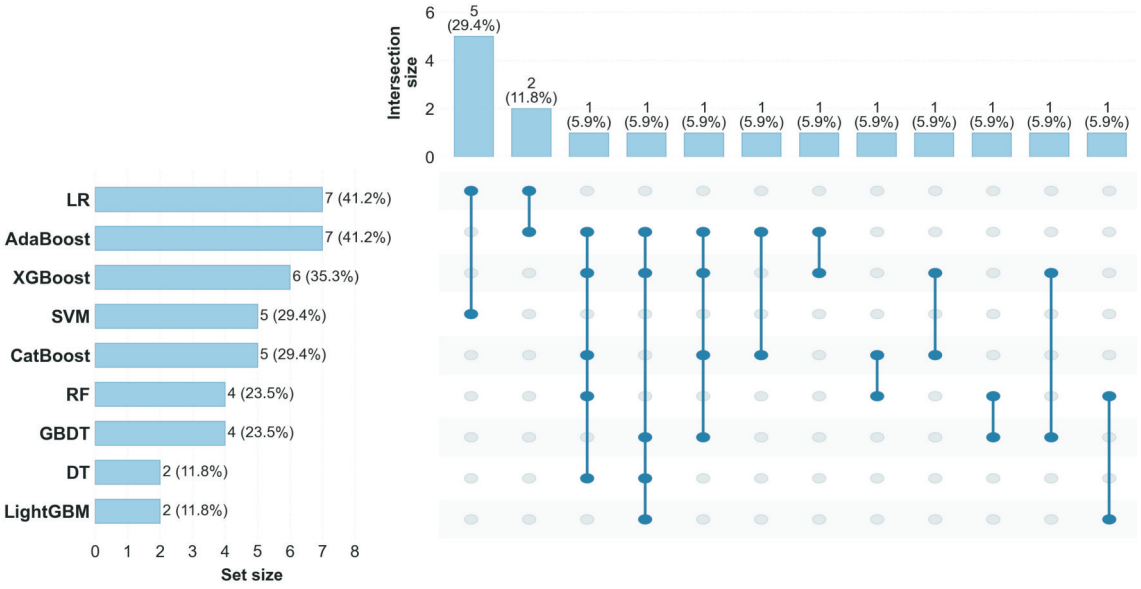

图3 基于多机器学习模型筛选的PF候选关键基因交集分析 UpSet 图

Figure 3. UpSet plot of intersection analysis of PF candidate key genes screened by multiple machine learning models

表3 PF候选关键基因的注释信息及功能描述

Table 3. Annotation information and functional description of PF candidate key genes

探针ID	基因符号	基因名称	基因功能描述
ILMN_21092	*POMT2*	蛋白质-*O*-甘露糖基转移酶2	催化特定蛋白质的*O*-甘露糖基化修饰，在维持蛋白质结构稳定性及功能调控中发挥重要作用
ILMN_17384	*UBE2Z*	泛素结合酶E2 Z	参与泛素化过程，将泛素分子转移至靶蛋白，从而介导其经蛋白酶体途径降解
ILMN_39439	*LOC654052*	泛素结合酶E2 Z假基因1	参与基因表达调控过程，其具体生物学功能尚不明确
ILMN_6746	*CCL18*	C-C模式趋化因子配体18	参与免疫调控过程，在炎症反应及纤维化进程中发挥作用，并可作为相关疾病的潜在生物标志物
ILMN_5441	*TXNL4B*	硫氧还蛋白样4B	参与pre-mRNA剪接过程，在RNA加工及基因表达调控中发挥关键作用

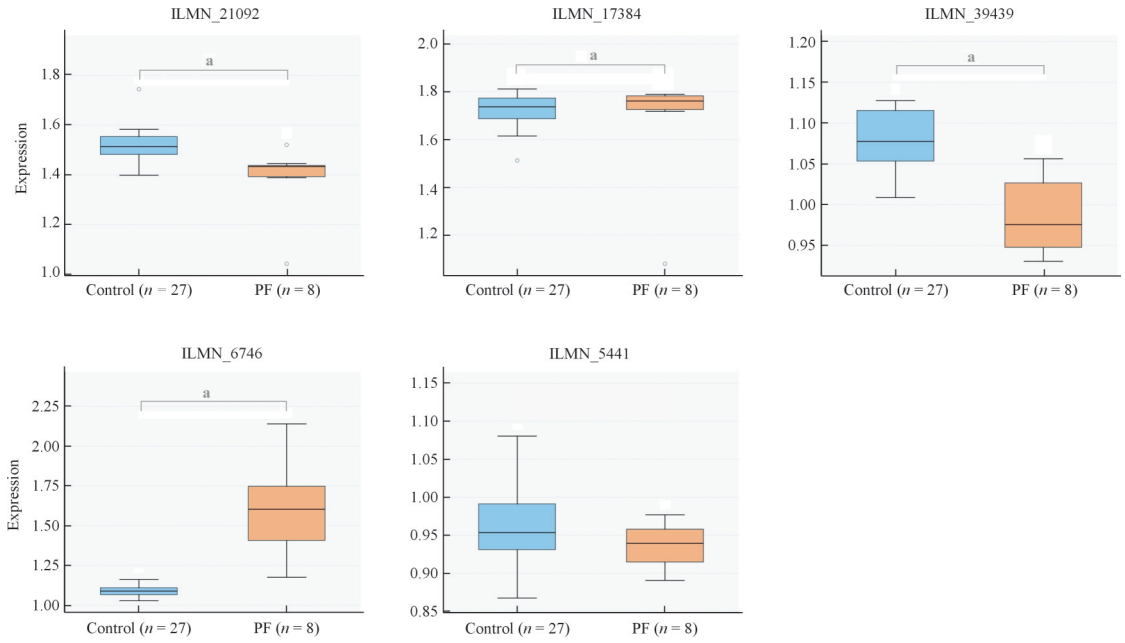

图4 PF候选关键基因在独立验证数据集GSE48149中的表达差异

Figure 4. Differential expression of PF core genes in the independent validation dataset GSE48149

注：$^{a}P < 0.05$。

为进一步辅助验证筛选结果的可靠性，本研究引入LASSO回归作为独立特征选择方法进行对照分析。在独立验证数据集GSE48149上通过交叉验证确定最优正则化参数λ，并筛选非零系数对应的特征作为候选基因集合。结果表明，*TXNL4B*、*CCL18*、*UBE2Z*等核心基因在LASSO方法中同样被稳定识别，表明多模型交集策略与基于正则化的特征选择方法具有良好一致性。

2.3 KEGG信号通路分析

KEGG分析结果显示，候选关键基因显著富集于多条与PF发生发展密切相关的通路。其中，甘露糖型*O*-糖基化生物合成（mannose type *O*-glycan biosynthesis）及其他类型*O*-糖基化生物合成（other types of *O*-glycan biosynthesis）通路的显著性最高（-lgFDR > 1.45），提示糖基化修饰过程在关键基因功能中占据重要地位。该结果与*POMT2*等参与蛋白糖基化修饰的基因功能相一致，表明ECM相关蛋白的翻译后修饰可能在PF过程中发挥重要作用。此外，炎症与免疫调控相关通路同样表现出显著富集，包括细胞因子-受体相互作用（cytokine-cytokine receptor interaction）、趋化因子信号通路（chemokine signaling pathway）以及病毒蛋白与细胞因子及其受体相互作用（viral protein interaction with cytokine and cytokine receptor）等。这些通路在调控炎症反应、免疫细胞募集及细胞间信号传递中发挥关键作用，提示筛选基因可能通过影响免疫微环境参与PF的发生发展，其中*CCL18*作为典型趋化因子在上述通路中具有重要调控作用。同时，泛素介导的蛋白降解通路（ubiquitin mediated proteolysis）显著富集，提示蛋白质稳态调控在PF过程中具有重要作用。该结果与*UBE2Z*等泛素结合酶相关基因的功能相吻合，表明异常蛋白降解及细胞应激反应可能参与疾病进展。综合来看，筛选得到的候选关键基因主要富集于糖基化修饰、炎症免疫调控及蛋白降解等关键生物通路，这些过程分别对应ECM重塑、免疫微环境失衡及细胞稳态调控等PF核心病理机制（图5）。

2.4 SHAP分析结果

SHAP摘要图清晰地揭示了各特征（基因）对模型输出的影响方向、强度及其与原始特征值的关系。*TXNL4B*的高表达（表现为红色）普遍对应着正向的SHAP值，表明其对模型输出具有显著的促进作用；相反，*UBE2Z*的高表达倾向于产生负向的SHAP值，提示其可能扮演抑制性角色。*CCL18*的影响具有双向性，即其表达水平在不同样本中既可增强也可减弱模型的预测倾向（图6），反映其在肺部炎症及免疫调控中的复杂作用。为进一步量化各特征的全局重要性，本研究计算了每个特征的平均绝对SHAP值，并绘制了特征重要性排序图，分析结果与UpSet摘要图高度一致，确认*TXNL4B*为最重要的预测因子，其平均绝对SHAP值最高，其次是*CCL18*和*UBE2Z*。而*LOC654052*和*POMT2*对模型整体预测的贡献较小（图7）。为进一步分析候选关键基因表达水平与模型预测结果之间的关系，绘制了SHAP依赖图，结果显示，不同基因在不同表达水平下对模型输出的影响呈现出一定的非线性变化趋势。*TXNL4B*在高表达水平时对应较高的SHAP值，表明其对疾病分类具有正向促进作用；而*UBE2Z*在高表达区域对应较低的SHAP值，提示其可能在模型判别中起到负向调控作用。*CCL18*的SHAP值分布则呈现一定离散性，说明其对模型输出的影响

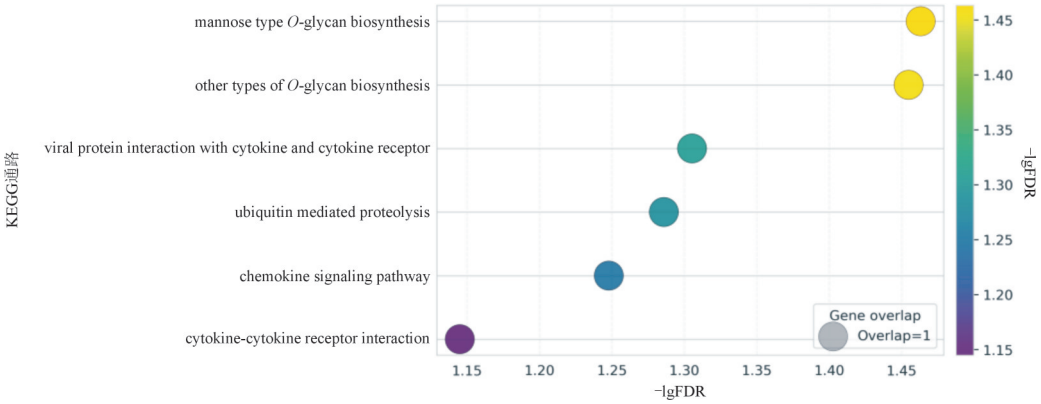

图5 基于GSE48149数据集的PF关键基因KEGG通路富集分析气泡图

Figure 5. Bubble chart of KEGG pathway enrichment analysis of core genes in PF based on the GSE48149 dataset

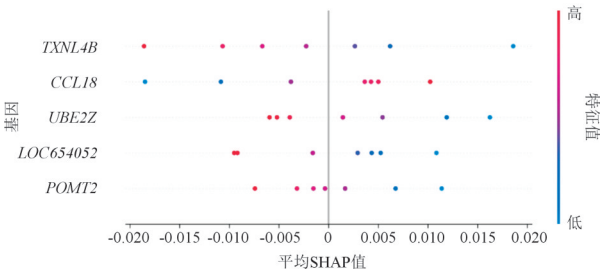

图6 PF候选关键基因的SHAP值分布

Figure 6. SHAP value distribution of PF candidate key genes

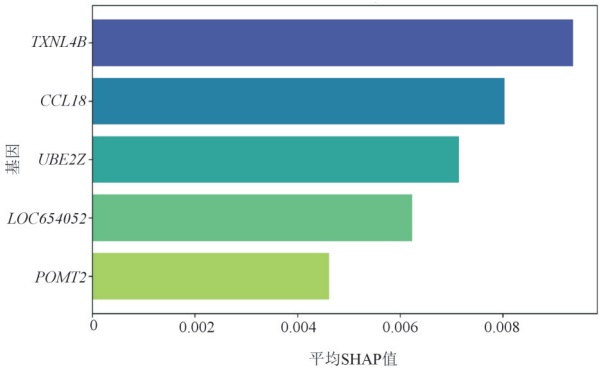

图7 SHAP 特征重要性

Figure 7. SHAP feature importance

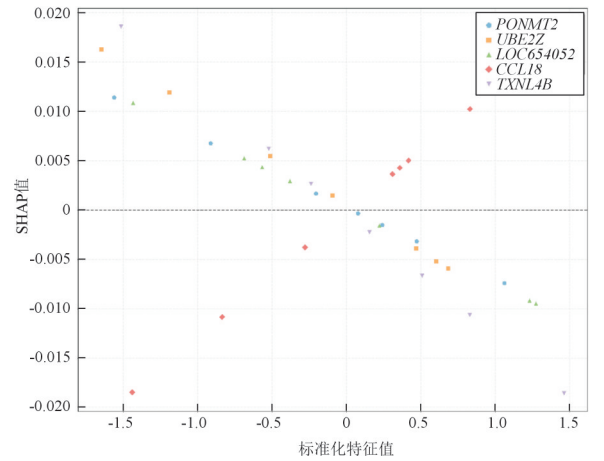

图8 基于SHAP的5个候选关键基因特征贡献依赖分析图

Figure 8. Contribution dependency analysis of 5 candidate key gene features based on SHAP

具有一定的复杂性，可能受到其他特征的协同调控。此外，*POMT2* 及 *LOC654052* 在不同表达区间的SHAP值变化相对平缓，表明其对模型预测结果的贡献较为稳定（图8）。

3 讨论

本研究构建了一种基于Stacking策略的集成学习模型，用于PF候选关键基因的筛选与分类识别。结果表明，该模型在训练数据集GSE70866的分层5折交叉验证中具有较好的分类性能（F1-score=0.9559，AUC=0.948 2），并在独立验证数据集GSE48149上表现出良好的泛化能力（召回率0.875，F1-score=0.933）。相较于单一模型在稳定性与预测准确性方面表现出一定优势，说明集成学习策略在高维小样本基因表达数据分析中具有良好的应用潜力。在特征筛选结果方面，通过多模型特征重要性综合分析，识别出 *POMT2*、*UBE2Z*、*LOC654052*、*CCL18* 和 *TXNL4B* 等关键基因。这些基因在炎症反应、免疫调控及ECM重塑等相关通路中呈现出一定的富集趋势，提示其可能参与PF相关分子过程。

可解释性方面，本研究通过元学习器权重和特征贡献度分析识别出 *TP53*、*FURIN*、*PON2* 等在既往研究中与PF相关的基因，同时筛选出 *POMT2*、*UBE2Z* 和 *CCL18* 等潜在新靶点，体现出模型在“复现经典—挖掘新标志物”之间取得的良好平衡。差异分析所筛选出的基因与模型的重要基因结果具有一定程度的重叠和互补性，其中多次被不同模型选出的候选关键基因（如 *POMT2*、*UBE2Z*、*CCL18*）可能在肺组织病理过程中具有更为关键的作用。相关研究表明，*POMT2* 编码的糖基转移酶在ECM蛋白的糖基化修饰中发挥重要作用，参与ECM蛋白的糖基化修饰[12]。*UBE2Z* 属于E2泛素结合酶家族，能够介导蛋白质的泛素化降解，可能通过影响蛋白降解及细胞应激反应参与PF相关调控[13]。*LOC654052* 虽为假基因，但越来越多的研究发现假基因可通过竞争性内源RNA机制调节miRNA活性，间接影响关键基因的转录表达[14]。另一方面，*CCL18* 作为分泌型趋化因子，在PF相关样本中呈现上调表达，并被报道与疾病进展相关，可作为纤维化进展的潜在生物标志物[15-16]。*TXNL4B* 则与RNA剪接复合体相关[17]，参与pre-mRNA加工过程，可能通过调控与细胞周期和应激反应相关的可变剪接事件影响纤维化过程，其在PF中的具体作用仍有待进一步研究。从生物学机制角度来看，这些基因分别涉及细胞凋亡调控、蛋白加工、线粒体氧化应激及ECM重塑等关键过程，与已有关于转化生长因子β调控成纤维细胞的研究结果一致，如转化生长因子β诱导的成纤维细胞激活与剪接调控[18]。此外，仅使用

模型筛选出的候选关键基因子集仍能达到接近完整模型的预测性能，进一步证明多基学习器输出的元特征融合提升模型性能。

从数据层面上，本研究主要基于GEO数据库中的公开转录组数据集（GSE70866与GSE48149）开展分析。尽管采用了训练集与独立验证集相结合的策略，但整体样本规模仍相对有限，且数据类型均来源于肺组织整体转录组，未能充分覆盖不同临床亚型及疾病进展阶段的分子特征。鉴于PF在临床表型及分子机制上具有显著的异质性，当前模型的泛化能力仍需在更大规模、多中心及多队列数据中进一步验证。其次，在方法层面，本研究主要依赖机器学习模型进行特征筛选与重要性评估，尽管通过多模型交集策略及LASSO回归方法进行一定程度的交叉验证，但仍缺乏对关键基因筛选结果的多维度验证。此外，本研究基于整体转录组数据进行分析，未能从细胞亚群水平刻画疾病微环境变化。PF的发生发展涉及成纤维细胞异常活化、免疫细胞浸润改变及ECM重塑等多层次调控过程[1]。未来可结合单细胞转录组及空间组学技术，从细胞类型特异性表达与组织空间分布角度进一步解析关键基因的作用机制，从而提升对疾病异质性的认识。

需要指出的是，本研究结果主要基于生物信息学分析与模型推断，上述候选关键基因的具体生物学功能及其潜在临床应用价值仍有待通过独立样本验证及体内外实验进一步确认。总体而言，本研究为高维组学数据条件下的关键基因筛选提供了一种可行的分析框架，但仍需在数据规模扩展与生物学功能验证方面持续深化。后续研究需通过结合多层次组学数据与临床信息，进一步提升模型的稳健性与临床转化潜力。

参考文献

1 Wang JH, Li K, Hao D, et al. Pulmonary fibrosis: pathogenesis and therapeutic strategies[J]. MedComm, 2024, 5(10): e744. DOI: 10.1002/mco2.744.

2 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis[J]. BMC Bioinformatics, 2008, 9(1): 559. DOI: 10.1186/1471-2105-9-559.

3 Rezk SS, Selim KS. Metaheuristic-based ensemble learning: an extensive review of methods and applications[J]. Neural Comput Appl, 2024, 36(29): 17931-17959. DOI: 10.1007/s00521-024-10203-4.

4 Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics[J]. Nat Rev Genet, 2015, 16(6): 321-332. DOI: 10.1038/nrg3920.

5 Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository[J]. Nucleic Acids Res, 2002, 30(1): 207-210. DOI: 10.1093/nar/30.1.207.

6 Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets-update[J]. Nucleic Acids Res, 2012, 41(D1): D991-D996. DOI: 10.1093/nar/gks1193.

7 Meaney C, Wang X, Guan J, et al. Comparison of methods for tuning machine learning model hyper-parameters: with application to predicting high-need high-cost health care users[J]. BMC Med Res Methodol, 2025, 25(1): 134. DOI: 10.1186/s12874-025-02561-x.

8 Hua Y, Wang L, Nguyen V, et al. A deep learning approach for transgender and gender diverse patient identification in electronic health records[J]. J Biomed Inform, 2023, 147: 104507. DOI: 10.1016/j.jbi.2023.104507.

9 Pisner DA, Schnyer DM. Support vector machine[M]. Amsterdam: Elsevier, 2020: 101-121.

10 He W, Su X, Chen L, et al. Potential biomarkers and therapeutic targets of idiopathic pulmonary arterial hypertension[J]. Physiol Rep, 2022, 10(1): e15101. DOI: 10.14814/phy2.15101.

11 Dalmolin M, Azevedo KS, Souza LCD, et al. Feature selection in cancer classification: utilizing explainable artificial intelligence to uncover influential genes in machine learning models [J]. AI, 2024, 6(1): 2. DOI: 10.3390/ai6010002.

12 De Bernabé DB-V, Inamori KI, Yoshida-Moriguchi T, et al. Loss of α-dystroglycan laminin binding in epithelium-derived cancers is caused by silencing of large[J]. J Biol Chem, 2009, 284(17): 11279-11284. DOI: 10.1074/jbc.C900007200.

13 Schelp J, Monte D, Dewitte F, et al. Structure of UBE2Z enzyme provides functional insight into specificity in the FAT10 protein conjugation machinery[J]. J Biol Chem, 2016, 291(2): 630-639. DOI: 10.1074/jbc.M115.671545.

14 Poliseno L. Pseudogenes: Functions and Protocols[M]. Cham: Springer, 2021: 131-147.

15 Prasse A, Probst C, Bargagli E, et al. Serum CC-chemokine ligand 18 concentration predicts outcome in idiopathic pulmonary fibrosis[J]. Am J Respir Crit Care Med, 2009, 179(8): 717-723. DOI: 10.1164/rccm.200808-1201OC.

16 Ghanbar MI, Villabona-Rueda A, Philip N, et al. Macrophage CCL18 promotes lung inflammation in checkpoint inhibitor pneumonitis[J]. Am J Respir Cell Mol Biol, 2025, Online ahead of print. DOI: 10.1165/rcmb.2025-0405OC.

17 Ju Z, Xiang J, Xiao L, et al. TXNL4B regulates radioresistance by controlling the PRP3-mediated alternative splicing of FANCI[J]. MedComm, 2023, 4(3): e258. DOI: 10.1002/mco2.258.

18 张正轩. TGF-β因子对原代成纤维细胞基因表达与可变剪接的调控机制研究[D]. 内蒙古包头：内蒙古科技大学, 2025. DOI: 10.27724/d.cnki.gnmgk.2025.000582.

收稿日期：2026年03月19日 修回日期：2026年05月18日
本文编辑：李 阳 冼静怡